

AMENDMENTS TO THE CLAIMS:

This listing of the claims will replace all prior versions, and listings, of the claims in this application.

Listing of Claims:

1. (Currently Amended) A method to process a text document, comprising:

partitioning text of the text document ~~text~~ and assigning semantic meaning to words of the partitioned ~~document~~ text, where assigning comprises applying a plurality of regular expressions, rules and a plurality of dictionaries to recognize chemical name fragments;

recognizing any substructures present in the chemical name fragments;

determining structural connectivity information of the chemical name fragments and recognized substructures;

extracting identifying information from the recognized chemical name fragments and substructures of the text document; and

storing the extracted identifying information in association with the determined structural connectivity information in a searchable index.
2. (Currently Amended) A method as in claim 1, wherein the extracting further comprises extracting keywords from the text document and wherein storing comprises storing the extracted

identifying information and the extracted keywords in association with the determined structural connectivity information in the searchable index, the method further comprising searching the index by a keyword and at least one of fragment name and substructure name.

3. (Currently Amended) A method as in claim 1, wherein extracting further comprises extracting keywords from the text document and wherein storing comprises storing the extracted identifying information and the extracted keywords in association with the determined structural connectivity information in the searchable index, the method further comprising searching the index by a keyword and at least one of fragment connectivity and substructure connectivity.

4. (Original) A method as in claim 1, further comprising searching the index by a combination of at least one of fragment and substructure name, and at least one of fragment and substructure connectivity.

5. (Original) A method as in claim 1, further comprising searching the index by at least one of fragment and substructure connectivity using a graphical user interface.

6. (Currently Amended) A method as in claim 1, wherein extracting further comprises extracting keywords from the text document and wherein storing comprises storing the extracted identifying information and the extracted keywords in association with the determined structural connectivity information in the searchable index, where the determined structural connectivity information is stored in a searchable structure index and the extracted keywords are stored in a text index,

further comprising searching the text index using at least one of a keyword, a fragment name and a substructure name and searching the structure index by at least one of fragment connectivity and substructure connectivity, and at an intersection of the search results from the structure index and the text index, identifying at least one document that contains a reference to a corresponding chemical compound.

7. (Original) A method as in claim 1, where determining structural connectivity information comprises looking up recognized fragments and substructures in a structure dictionary.

8. (Original) A method as in claim 7, where the structure dictionary comprises at least one of a MOL dictionary and a SMILES dictionary.

9. (Original) A method as in claim 1, where said plurality of dictionaries comprise a dictionary of common chemical prefixes and a dictionary of common chemical suffixes.

10. (Original) A method as in claim 1, where said plurality of dictionaries comprise a dictionary of stop words to eliminate erroneous chemical name fragments.

11. (Original) A method as in claim 1, further comprising filtering recognized chemical name fragments using a list of stop words to eliminate erroneous chemical name fragments.

12. (Original) A method as in claim 1, where chemical name fragments are further recognized by

using common chemical word endings.

13. (Original) A method as in claim 1, where application of said regular expressions and rules results in punctuation characters being one of maintained or removed between chemical name fragments as a function of context.

14. (Original) A method as in claim 1, where said regular expressions comprise a plurality of patterns, individual ones of which are comprised of at least one of characters, numbers and punctuation.

15. (Original) A method as in claim 14, where the punctuation comprises at least one of parenthesis, square bracket, hyphen, colon and semi-colon.

16. (Original) A method as in claim 14, where the characters comprise at least one of upper case C, O, R, N and H.

17. (Original) A method as in claim 14, where the characters comprise strings of at least one of lower case xy, ene, ine, yl, ane and oic.

18. (Original) A method as in claim 1, comprising an initial step of tokenizing the document to provide a sequence of tokens.

19. (Currently Amended) A system to process a text document, comprising:

a unit to partition text of the text document ~~text~~ and to assign semantic meaning to words of the partitioned ~~document~~ text, where assigning comprises applying a plurality of regular expressions, rules and a plurality of dictionaries to recognize chemical name fragments;

a unit to recognize any substructures present in the chemical name fragments;

a unit to extract identifying information from the recognized chemical name fragments and substructures of the text document; and

a unit to determine structural connectivity information of the chemical name fragments and recognized substructures and to store the extracted identifying information in association with the determined structural connectivity information in a searchable index.

20. (Currently Amended) A system as in claim 19, wherein the unit to extract is further to extract keywords from the text document and wherein the unit to store is further to store the extracted identifying information and the extracted keywords in association with the determined structural connectivity information in the searchable index, the system further comprising a unit to search the index by a keyword and at least one of fragment name and substructure name.

21. (Currently Amended) A system as in claim 19, wherein the unit to extract is further to extract keywords from the text document and wherein the unit to store is further to store the extracted

identifying information and the extracted keywords in association with the determined structural connectivity information in the searchable index, the system further comprising a unit to search the index by a keyword and at least one of fragment connectivity and substructure connectivity.

22. (Original) A system as in claim 19, further comprising a unit to search the index by a combination of at least one of fragment and substructure name, and at least one of fragment and substructure connectivity.

23. (Original) A system as in claim 19, further comprising a unit to search the index by at least one of fragment and substructure connectivity using a graphical user interface.

24. (Currently Amended) A system as in claim 19, wherein the unit to extract is further to extract keywords from the text document and wherein the unit to store is further to store the extracted identifying information and the extracted keywords in association with the determined structural connectivity information in the searchable index, where the determined structural connectivity information is stored in a searchable structure index and the extracted keywords are stored in a text index, the system further comprising a unit to search the text index using at least one of a keyword, a fragment name and a substructure name and to search the structure index by at least one of fragment connectivity and substructure connectivity, and at an intersection of the search results from the structure index and the text index, to identify at least one document that contains a reference to a corresponding chemical compound.

25. (Original) A system as in claim 19, where said unit that determines structural connectivity information looks up recognized fragments and substructures in a structure dictionary.

26. (Original) A system as in claim 25, where the structure dictionary comprises at least one of a MOL dictionary and a SMILES dictionary.

27. (Original) A system as in claim 19, where said plurality of dictionaries comprise a dictionary of common chemical prefixes and a dictionary of common chemical suffixes.

28. (Original) A system as in claim 19, where said plurality of dictionaries comprise a dictionary of stop words to eliminate erroneous chemical name fragments.

29. (Original) A system as in claim 19, further comprising a unit to filter recognized chemical name fragments using a list of stop words to eliminate erroneous chemical name fragments.

30. (Original) A system as in claim 19, where chemical name fragments are further recognized by using common chemical word endings.

31. (Original) A system as in claim 19, where application of said regular expressions and rules results in punctuation characters being one of maintained or removed between chemical name fragments as a function of context.

32. (Original) A system as in claim 19, where said regular expressions comprise a plurality of patterns, individual ones of which are comprised of at least one of characters, numbers and punctuation.

33. (Original) A system as in claim 32, where the punctuation comprises at least one of parenthesis, square bracket, hyphen, colon and semi-colon.

34. (Original) A system as in claim 32, where the characters comprise at least one of upper case C, O, R, N and H.

35. (Original) A system as in claim 32, where the characters comprise strings of at least one of lower case xy, ene, ine, yl, ane and oic.

36. (Original) A system as in claim 19, further comprising an input tokenizer unit to receive documents to be processed to provide a sequence of tokens.

37. (Currently Amended) A computer program product for storing in a computer readable form a set of computer program instructions for directing at least one computer to process text of a text document, comprising instructions to parse the text of the text document ~~text~~ to recognize chemical name fragments; instructions to recognize any substructures present in the chemical name fragments; instructions to extract identifying information from the recognized chemical name fragments and substructures of the text document; and instructions to determine structural

connectivity information of the chemical name fragments and recognized substructures and to store the extracted identifying information in association with the determined structural connectivity information in a searchable index.

38. (Currently Amended) A computer program product as in claim 37, wherein the instructions to extract identifying information further extract keywords from the text document and wherein the instructions to store further store the extracted identifying information and the extracted keywords in association with the determined structural connectivity information in the searchable index, the program further comprising instructions to search the index by a keyword and at least one of fragment name and substructure name.

39. (Currently Amended) A computer program product as in claim 37, wherein the instructions to extract identifying information further extract keywords from the text document and wherein the instructions to store further store the extracted identifying information and the extracted keywords in association with the determined structural connectivity information in the searchable index, the program further comprising instructions to search the index by a keyword and at least one of fragment connectivity and substructure connectivity.

40. (Original) A computer program product as in claim 37, further comprising instructions to search the index by a combination of at least one of fragment and substructure name, and at least one of fragment and substructure connectivity.

41. (Original) A computer program product as in claim 37, further comprising instructions to search the index by at least one of fragment and substructure connectivity using a graphical user interface.

42. (Currently Amended) A computer program product as in claim 37, wherein the instructions to extract identifying information further extract keywords from the text document and wherein the instructions to store further store the extracted identifying information and the extracted keywords in association with the determined structural connectivity information in the searchable index, where the determined structural connectivity information is stored in a searchable structure index and the extracted keywords are stored in a text index, and further comprising instructions to search the text index using at least one of a keyword, a fragment name and a substructure name and to search the structure index by at least one of fragment connectivity and substructure connectivity, and at an intersection of the search results from the structure index and the text index, to identify at least one document that contains a reference to a corresponding chemical compound.

43. (Currently Amended) A system comprising a plurality of computers at least two of which are coupled together through a data communications network, said system comprising a unit to parse text of a text document text to recognize chemical name fragments; a unit to recognize any substructures present in the chemical name fragments; a unit to extract identifying information of the recognized chemical name fragments and substructures from the text document; and a unit to determine structural connectivity information of the chemical name fragments and recognized

substructures and to store the extracted identifying information in association with the determined structural connectivity information in a searchable index.

44. (Currently Amended) A system as in claim 43, wherein the unit to extract is further to extract keywords from the text document and wherein the unit to store is further to store the extracted identifying information and the extracted keywords in association with the determined structural connectivity information in the searchable index, where the determined structural connectivity information is stored in a searchable structure index and the extracted keywords are stored in a text index, the system further comprising a unit to search the text index using at least one of a keyword, a fragment name and a substructure name and to search the structure index by at least one of fragment connectivity and substructure connectivity, and at an intersection of the search results from the structure index and the text index, to identify at least one document that contains a reference to a corresponding chemical compound.

45. (Original) A system as in claim 43, where said unit that determines structural connectivity information looks up recognized fragments and substructures in a structure dictionary.

46. (Original) A system as in claim 45, where the structure dictionary comprises at least one of a MOL dictionary and a SMILES dictionary.